Towards perceptually optimized sound zones

A proof-of-concept study

Lee, Taewoong; Nielsen, Jesper Kjær; Christensen, Mads Græsbøll

Published in:
2019 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2019 - Proceedings

DOI (link to publication from Publisher):
10.1109/ICASSP.2019.8682902

Publication date:
2019

Document Version
Accepted author manuscript, peer reviewed version

TOWARDS PERCEPTUALLY OPTIMIZED SOUND ZONES: A PROOF-OF-CONCEPT STUDY

Taewoong Lee, Jesper Kjær Nielsen, and Mads Græsbøll Christensen

Audio Analysis Lab, CREATE, Aalborg University, Aalborg, Denmark
{tlee, jkn, mgc}@create.aau.dk

ABSTRACT

The creation of sound zones has been an active research topic for approximately two decades. Many sound zone control methods have been proposed, and the best approaches result in a target to interferer ratio (TIR) of about 15 dB in a practical set-up. Unfortunately, this is far from a TIR of about 25 dB which is currently believed necessary to make sound zones commercially viable. However, state-of-the-art sound zone control methods take neither the input signal characteristics nor human auditory perception into account. In this paper, we show how a recently proposed sound zone control framework called VAST can be extended into perceptual VAST (P-VAST) which takes input signal characteristics and human auditory perception into account. We also make a proof-of-concept simulation and an AB preference test which both show that P-VAST outperforms traditional sound zone control methods in terms of perceptually meaningful metrics such as STOI and PESQ in a fairly simple set-up.

Index Terms— Sound zones, human auditory system, masking effect, variable span trade-off filter, personal sound

1. INTRODUCTION

Sound zones allow multiple people in the same acoustical space to enjoy their own desired audio contents without using headphones. Generally, two types of zones are considered: a bright zone and a dark zone. The bright zone is a confined region in which a desirable sound field is reproduced as faithfully as possible, whereas the dark zone is a confined region in which the sound field is suppressed as much as possible. These two zones are created by controlling a loudspeaker array, and multiple bright zones can be created by superposing the individual bright and dark zone solutions for every input signal. Sound zones have many applications including outdoor concert [1], mobile phones [2], car cabins [3], personal computer [4], and other applications [5, 6].

The performance of sound zone control strategies is typically measured in terms of either the acoustic contrast, the reproduction error, or the target to interferer ratio (TIR). The acoustic contrast describes the ratio of the acoustic potential energy between the bright and dark zones, and this is maximized in the control strategy known as acoustic contrast control (ACC) [7]. On the other hand, the reproduction error, i.e., signal distortion, describes the difference between the reproduced and desired sound fields, and this is minimized in the control strategy known as pressure matching (PM) [8]. Finally, TIR is a zone-specific measure describing the ratio of either the acoustic potential energy or loudness between the desired and interfering sound fields in a given zone (see [9] for more on this). Unfortunately, a high contrast and a low distortion cannot be optimized for simultaneously. Instead, ACC maximizes the acoustic contrast and the signal distortion, whereas PM minimizes both. This has been recently shown in [10] where a framework referred to as variable span trade-off filter (VAST) was proposed. VAST allows one to trade-off the acoustic contrast for the signal distortion and vice versa.

To make sound zones commercially viable, the perceived separation between the desired and interfering sound fields in a zone should be high enough, and the reproduction error should be small enough. A recent listening experiment in [11] on both speech and music signals found that at least 25 dB of TIR (loudness-based) is needed to give a distraction score of at most 10 (out of 100) where the distraction is defined as how much the interfering signal is taking the attention away from the desired signal [12]. Unfortunately, BACC-PM [13, 14], which was recently proposed and used as the sound zone control method in the above study, has only been reported to give a TIR of around 15 dB in a practical set-up. On the surface, this rather large gap between the reported and needed TIR seems discouraging for the applicability of the sound zone technology, but it is important to stress that the needed TIR should be seen in the context of the control method being used rather than a universal design criterion that all control methods should fulfill.

To the best of our knowledge, except [15] where the noise masker has been introduced to hide the reproduced speech in the dark zone and preserve the quality of that in the bright zone, existing sound zone control methods such as [7, 8, 13, 14, 16–19] neither take the input signal characteristics nor human auditory perception into account. Instead, the control filters are typically found for white input signals and physically meaningful metrics such as the mean squared error and acoustic potential energy. The main advantage of this approach is that the control filters can be calculated offline using convex optimization methods, but this comes at the cost of wasted control efforts on controlling frequencies which are perhaps not present in the input signal or inaudible. For audio coding, a famous demonstration at the AT&T Bell Labs in the early 1990s showed that exploiting the input signal characteristics and human auditory perception dramatically lowered the signal-to-quantization-noise ratio required for perceptually transparent audio coding. In particular, the demonstration, which is now often referred to as the "13dB miracle" [20, 21], showed that the quantization noise corresponding to a segment-wise SNR of 13 dB is inaudible if the noise spectrum is shaped according to the human auditory system. Based on this observation, the underlying hypothesis of our work is that the requirements to the TIR can be dramatically lowered if the input signal characteristic and the human auditory perception are taken into account, and we here take a first step to confirm this hypothesis.

In this paper, we propose a new framework called perceptual VAST (P-VAST) for creating sound zones which takes the input signal characteristics and the human auditory system into account. As the name suggests, P-VAST is an extension of our recently proposed VAST framework [10] which has many existing sound zone control methods as special cases. Via simulations and a listening test on a proof-of-concept implementation, we show that we can significantly improve on perceptual metrics without improving physical metrics.

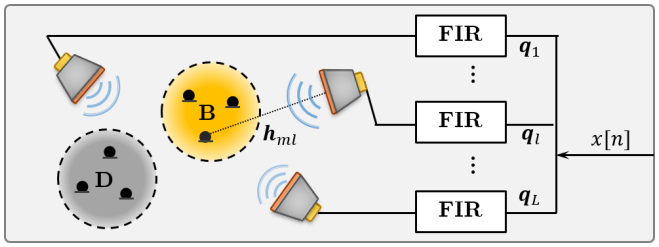

Fig. 1. The system setup of sound zones

2. WEIGHTED VAST FRAMEWORK

In this section, we describe the VAST framework from [10] and extend it with an arbitrary weighting on the reproduction error. This will set us up for the introduction of P-VAST in Sec. 3. To derive VAST, we initially consider the simple illustration in Fig. 1. In the figure, the bright and dark zones are spatially sampled by M_B and M_D microphone positions, respectively[1]. At the mth microphone position or, equivalently, control point, the reproduced sound pressure is a convolution between the input signal $x[n]$, the L J-dimensional control filters $\{\boldsymbol{q}_l\}_{l=1}^{L}$, and the L K-dimensional room impulse responses $\{\boldsymbol{h}_{ml}\}_{l=1}^{L}$, i.e.,

$$\begin{aligned} p_m[n] &= \sum_{l=1}^{L}\sum_{k=0}^{K-1}\sum_{j=0}^{J-1} x[n-k-j]h_{ml}[k]q_l[j] \\ &= \sum_{l=1}^{L} \boldsymbol{y}_{ml}^T[n]\boldsymbol{q}_l = \boldsymbol{y}_m^T[n]\boldsymbol{q} \end{aligned} \tag{1}$$

where we have defined

$$\boldsymbol{y}_{ml}[n] = \boldsymbol{X}[n]\boldsymbol{h}_{ml} \tag{2}$$

$$\boldsymbol{X}[n] = \begin{bmatrix} x[n] & \cdots & x[n-K+1] \\ \vdots & \ddots & \vdots \\ x[n-J+1] & \cdots & x[n-K-J+2] \end{bmatrix} \tag{3}$$

$$\boldsymbol{y}_m[n] = \begin{bmatrix} \boldsymbol{y}_{m1}^T[n] & \cdots & \boldsymbol{y}_{mL}^T[n] \end{bmatrix}^T \tag{4}$$

$$\boldsymbol{q} = \begin{bmatrix} \boldsymbol{q}_1^T & \cdots & \boldsymbol{q}_L^T \end{bmatrix}^T . \tag{5}$$

Note that $\boldsymbol{y}_{ml}[n]$ can be interpreted as the uncontrolled reproduced pressure at microphone position m originating from loudspeaker l.

Sound zone control is about designing the control filters in $\boldsymbol{q}$ so that the desired sound pressures at the M_B and M_D control points are reproduced as faithfully as possible. For the control points in the dark zone, the desired pressures are all 0 whereas the desired sound pressures in the bright zone are here defined as a sound field generated by a virtual source at a point z emitting $x[n]$. Thus, the desired sound pressure at the mth control point is

$$d_m[n] = \begin{cases} (h_{mz} * x)[n] & m \in \mathcal{M}_B \\ 0 & m \in \mathcal{M}_D \end{cases}, \tag{6}$$

where $\mathcal{M}_B$ and $\mathcal{M}_D$ are the set of microphone indices for the bright and dark zones, respectively, and $h_{mz}[n]$ is the impulse response from the virtual source to the mth control point. Note that if the desired signal is generated under assumed anechoic conditions, the sound zone control method essentially also has to perform dereverberation in order to match the desired and reproduced sound fields. The reproduction error at the mth microphone is traditionally defined as the difference between the desired sound pressure and the reproduced sound pressure, i.e.,

$$\varepsilon_m[n] = d_m[n] - p_m[n] . \tag{7}$$

However, we will here consider the more general weighted reproduction error defined as

$$\tilde{\varepsilon}_m[n] = (w_m * \varepsilon_m)[n] = \tilde{d}_m[n] - \tilde{p}_m[n] \tag{8}$$

where, e.g., $\tilde{p}_m[n]$ means that we have filtered $p_m[n]$ with the weighting filter $w_m[n]$, i.e.,

$$\tilde{p}_m[n] = (w_m * p_m)[n] = \sum_{l=1}^{L} \tilde{\boldsymbol{y}}_{ml}^T[n]\boldsymbol{q}_l = \tilde{\boldsymbol{y}}_m^T[n]\boldsymbol{q} . \tag{9}$$

Note that the weighting filter is assumed known and is used to shape the reproduction error according to some design criterion. We return to this in Sec. 3.

Based on the above definitions, we can now define the weighted signal distortion power $\tilde{S}_B(\boldsymbol{q})$ and the weighted residual error power $\tilde{S}_D(\boldsymbol{q})$ as

$$\begin{aligned} \tilde{S}_B(\boldsymbol{q}) &= \frac{1}{M_B N}\sum_{n=0}^{N-1}\sum_{m\in\mathcal{M}_B} |\tilde{\varepsilon}_m[n]|_2^2 \\ &= \tilde{\sigma}_d^2 - 2\boldsymbol{q}^T\tilde{\boldsymbol{r}}_B + \boldsymbol{q}^T\tilde{\boldsymbol{R}}_B\boldsymbol{q} \end{aligned} \tag{10a}$$

$$\tilde{S}_D(\boldsymbol{q}) = \frac{1}{M_D N}\sum_{n=0}^{N-1}\sum_{m\in\mathcal{M}_D} |\tilde{\varepsilon}_m[n]|_2^2 = \boldsymbol{q}^T\tilde{\boldsymbol{R}}_D\boldsymbol{q}, \tag{10b}$$

where N is the number of observations and

$$\tilde{\sigma}_d^2 = \frac{1}{M_B N}\sum_{n=0}^{N-1}\sum_{m\in\mathcal{M}_B} |\tilde{d}_m[n]|_2^2$$

$$\tilde{\boldsymbol{r}}_B = \frac{1}{M_B N}\sum_{n=0}^{N-1}\sum_{m\in\mathcal{M}_B} \tilde{\boldsymbol{y}}_m[n]\tilde{d}_m[n]$$

$$\tilde{\boldsymbol{R}}_C = \frac{1}{M_C N}\sum_{n=0}^{N-1}\sum_{m\in\mathcal{M}_C} \tilde{\boldsymbol{y}}_m[n]\tilde{\boldsymbol{y}}_m^T[n] \quad \text{for } C \in \{B, D\} .$$

We can now define an objective function as

$$J(\boldsymbol{q}) = \tilde{S}_B(\boldsymbol{q}) + \mu\tilde{S}_D(\boldsymbol{q}) \tag{11}$$

where $\mu \geq 0$ is a user-defined parameter which controls the relative importance of suppressing the residual error power. When the weighting filter is the Kronecker delta function, the solution to this optimization problem is known as the ACC-PM method [19]. As argued in [10], however, it is typically not advantageous to minimize this objective directly w.r.t. $\boldsymbol{q}$, but instead w.r.t. a low-rank approximation to $\boldsymbol{q}$. Specifically, assuming $\tilde{\boldsymbol{R}}_D$ has full rank, we first compute the (generalized) eigenvalue decomposition [22]

$$\tilde{\boldsymbol{R}}_D^{-1}\tilde{\boldsymbol{R}}_B\boldsymbol{U}_{LJ} = \boldsymbol{U}_{LJ}\boldsymbol{\Lambda}_{LJ} \tag{12}$$

where $\boldsymbol{\Lambda}_{LJ}$ is a diagonal matrix containing the eigenvalues in descending order and $\boldsymbol{U}_{LJ}$ is an invertible matrix containing the corresponding eigenvectors. We then form the low rank approximation to $\boldsymbol{q}$ as a linear combination of the first V eigenvectors, i.e.,

$$\boldsymbol{q} \approx \boldsymbol{U}_V\boldsymbol{a}_V \tag{13}$$

and optimize the quadratic objective $J(\boldsymbol{U}_V\boldsymbol{a}_V)$ w.r.t. $\boldsymbol{a}_V$. The solution to this optimization problem can be derived analytically and is given by

$$\boldsymbol{a}_{\text{VAST}} = \arg\min_{\boldsymbol{a}_V} J(\boldsymbol{U}_V\boldsymbol{a}_V) = \left[\boldsymbol{\Lambda}_V + \mu\boldsymbol{I}_V\right]^{-1}\boldsymbol{U}_V^T\tilde{\boldsymbol{r}}_B \tag{14}$$

[1]Throughout this paper, the subscripts $_B$ and $_D$ represent the bright and dark zones, respectively.

Table 1. Desired signal and masker for a control point m

Zone	$\alpha \quad (m \in \mathcal{M}_\alpha)$	$\beta \quad (m \in \mathcal{M}_\beta)$
Desired signal	$d_m^{(\alpha)}[n]$	$d_m^{(\beta)}[n]$
Masker	$d_m^{(\alpha)}[n]$	$d_m^{(\beta)}[n]$

so that

$$\boldsymbol{q}_{\text{VAST}}(V,\mu) = \boldsymbol{U}_V \boldsymbol{a}_{\text{VAST}}(V,\mu) = \sum_{v=1}^{V} \frac{\boldsymbol{u}_v^T \tilde{\boldsymbol{r}}_B}{\lambda_v + \mu} \boldsymbol{u}_v, \qquad (15)$$

where λ_v and $\boldsymbol{u}_v$ are the vth eigenvalue and eigenvector, respectively. Interestingly, we obtain the ACC and ACC-PM methods as special cases of VAST for $V = 1$ and $V = LJ$, respectively. For more on VAST and its special cases, we refer the interested reader to [10]. Here, we will now focus on how the weighting filter can be designed to take human auditory perception into account.

3. PERCEPTUAL VAST

As alluded to in the introduction, audio coding was revolutionized by doing the coding according to simple mathematical models for human auditory perception. The human auditory system has a limited time- and frequency resolution so a certain sound, the so-called maskee, becomes less or inaudible in the presence of a stronger masker close to the maskee in the time- and/or frequency domain [23]. This phenomenon is generally referred to as masking, and it allows us to make surprisingly large modifications to audio signals without changing how they are perceived. In some applications as, e.g., in the creation of sound zones, we have so many and often conflicting design constraints that it is impossible to render all signal modifications inaudible. In this case, however, we can still minimize the amount of distraction of the signal modifications by taking the human auditory system into account.

For simplicity, we here exclusively focus on taking into account spectral or simultaneous masking which is also much more significant than temporal or non-simultaneous masking[2]. The idea is straight-forward: For a given segment and control point m, we compute a masking curve based on the psychoacoustic model in [24]. This masking curve, which models the threshold below which all interfering sounds are inaudible, is then inverted and used as the weighting filter $w_m[n]$. In terms of the weighted error $\tilde{\varepsilon}_m[n]$, this means that we allow for bigger errors in those part of the spectrum where the masker has a high power and penalize errors in those part of the spectrum where the masker has a low or no power.

A fundamental question is, of course, what the masker is at a control point m. To answer this question, we initially consider the case of two zones labelled α and β, respectively. For a control point $m \in \mathcal{M}_\alpha$, i.e., a control point in zone α, we assume that the desired signal is $d_m^{(\alpha)}[n]$ whereas the desired signal for a control point $m \in \mathcal{M}_\beta$, i.e., a control point in zone β, is $d_m^{(\beta)}[n]$. Similarly, the input signals for the two zones are $x^{(\alpha)}[n]$ and $x^{(\beta)}[n]$, respectively. When we design the control filters for reproducing $d_m^{(\alpha)}[n]$ $\forall m \in \mathcal{M}_\alpha$, zone α acts as the bright zone and zone β acts as the dark zone. Thus, the masker for a control point $m \in \mathcal{M}_\alpha$ is $d_m^{(\alpha)}[n]$ whereas the masker for a control point $m \in \mathcal{M}_\beta$ is $d_m^{(\beta)}[n]$. Conversely, zone β acts as the bright zone and zone α as the dark zone when we design the control filters for reproducing $d_m^{(\beta)}[n]$ $\forall m \in \mathcal{M}_\beta$. Consequently, the masker for a control point $m \in \mathcal{M}_\beta$, which is now in the bright zone, is $d_m^{(\beta)}[n]$ whereas the masker for a control point $m \in \mathcal{M}_\alpha$, which is now in the dark zone, is $d_m^{(\alpha)}[n]$.

[2]We note in passing that when signals are processed on a segment-by-segment basis some sort of temporal masking is also indirectly exploited.

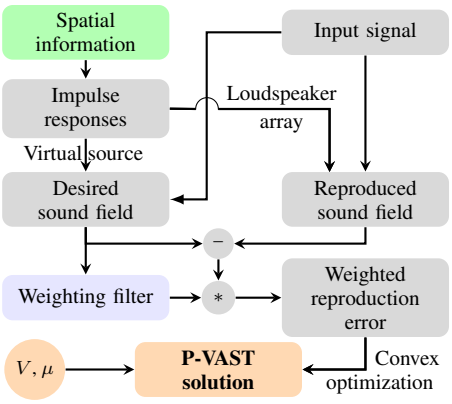

Fig. 2. The flowchart of P-VAST

In Table 1, we have summarized the above discussion. The essence is that the masker at a control point is the desired signal at this point, regardless of the number of zones and whether the point is in the bright or dark zone. Labelling a zone as bright or dark is namely nothing but a convenient abstraction which allows us to design the control filters of each zone individually. However, the masker does not change depending on which zone we are currently computing the control filters for. In the case where we wish to have silence in a zone, the desired signal for the control points in this zone are all zero, and the masking curves will simply be the threshold in quiet. When we design the weighting filters as described here, we obtain perceptual VAST (P-VAST) which is illustrated in Fig. 2.

4. EXPERIMENTS

In order to verify P-VAST, objective evaluations and an AB preference test were conducted. For the system as shown in Fig. 3 (a), a virtual source and a circular array with evenly distributed loudspeakers were considered, each zone was spatially sampled at 16 control points. We assumed that all loudspeakers and microphones were located in the same plane, that all loudspeakers behaved as point sources, and that no reverberation was present. The length of the control filters $\{\boldsymbol{q}_l\}_{l=1}^{L}$ was $J = 640$, the length of the impulse responses was $K = 512$, and the sampling frequency was 12.8 kHz which is commonly used in sound and vibration analysis. The parameters for P-VAST were chosen as $\mu_\alpha = 0.8$, $\mu_\beta = 0.7$ and $V_\alpha = V_\beta = LJ/4$. The selected audio contents were 6 seconds of conversations excerpted from the movie "Zootopia" by Disney in two different languages, English and Danish, which were set as the desired signals in zone α and β, respectively, and down-sampled from 44.1 kHz to 12.8 kHz. The energy of the two input signals were set to be the same. To avoid having to recompute the control filters for every short signal segment, we computed masking thresholds for segments of 200 ms and used the inverse of the average of these for computing the weighting filters. The spatial covariance matrices were also computed for the entire signal instead of on a per segment basis; thus, the control filters only had to be computed once for every signal which is desirable from a computational point of view but clearly suboptimal in terms of fully exploiting the masking effect. However, it turned out that this simple approach still resulted in a

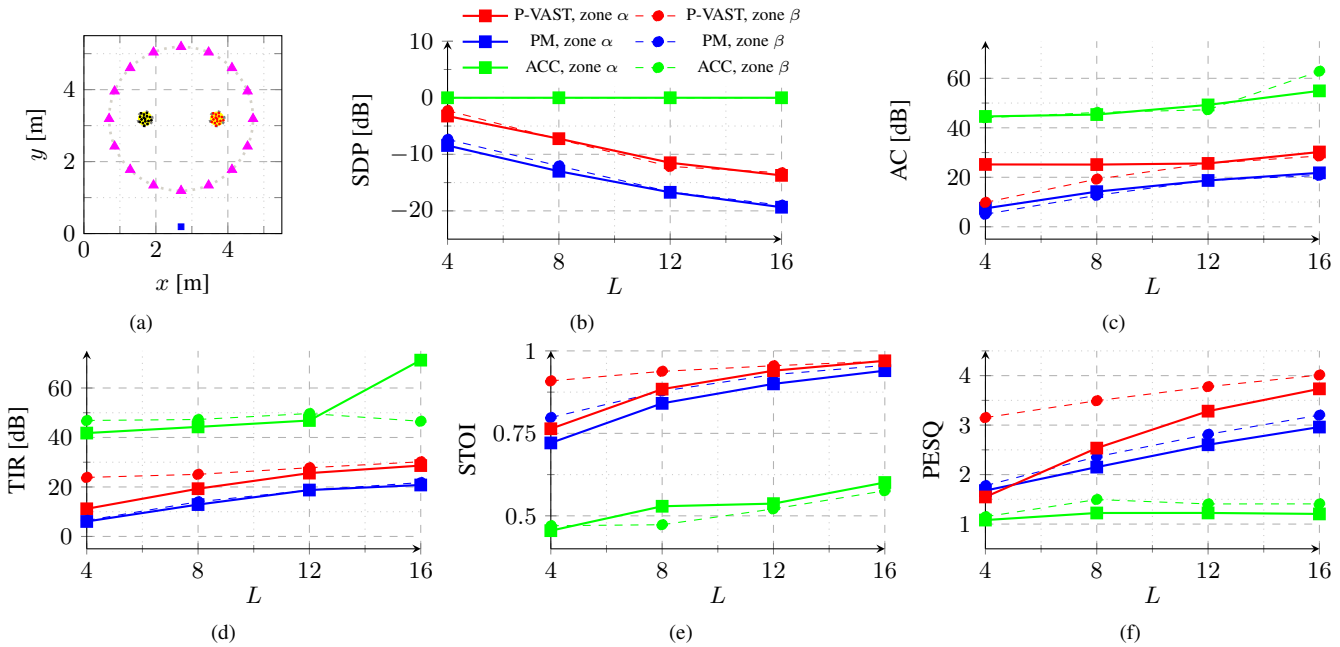

Fig. 3. (a) An example of the system geometry with 16 loudspeakers (▲), virtual source (■), and control points (●) which were sampled according to Vogel's method [25]. The centers of the zones with radii 0.2 m were 2 m apart from each other. (b) – (f) SDP, AC, TIR, STOI, and PESQ for each zone for various methods with respect to the number of loudspeakers.

Table 2. Results of AB preference test with a reference (S and M denote speech and music, respectively)

		Preference [%]					
Set		PM	P-VAST	μ_α	μ_β	V_α	V_β
1	S, S	14.7	85.3	0.5	0.3	$LJ/4$	$LJ/4$
2	M, S	26.5	73.5	0.5	0.4	$LJ/4$	$LJ/4$
3	M, M	5.9	94.1	0.3	0.1	$LJ/2$	$LJ/4$
4	S, S	10.3	89.7	0.8	0.7	$LJ/4$	$LJ/4$

large improvement over existing methods, and the results presented here should, therefore, be seen as a simple proof-of-concept.

In the first experiment, we compared P-VAST to the existing methods: ACC and PM. Since speech signals were used, we assessed the reproduced sound fields by using not only the signal distortion power (SDP) $S_B(\boldsymbol{q})$ in (10a), the acoustic contrast (AC) defined by $\frac{M_D}{M_B}(\boldsymbol{q}^T\boldsymbol{R}_B\boldsymbol{q})/(\boldsymbol{q}^T\boldsymbol{R}_D\boldsymbol{q})$, and TIR defined as the ratio of the acoustic potential energy between the desired and interfering sound fields, but also the short-time objective intelligibility (STOI) [26] and perceptual evaluation of speech quality (PESQ) [27]. All metrics were spatially averaged over all control points in each zone. Note that no weighting was used in the computation of reproduced fields. The results of the objective evaluations are shown in Fig. 3 (b) – (f).

A number of interesting observations can be made from Fig. 3 (b) – (f). First, and unsurprisingly, all metrics improve with an increasing number of loudspeakers. Second, for the physical measures (SDP, AC, and TIR), ACC has the highest values, PM the lowest, and P-VAST something in between these two extrema, something which for SDP and AC can also be shown to hold theoretically in general [10]. Third, the perceptual measures STOI and PESQ show that P-VAST significantly outperforms ACC and PM. Although we have only used average masking curves and signal statistics, this has clearly demonstrated the large potential of taking the input signal characteristics and human auditory perception into account. To further support this claim, an AB preference test with reference was conducted on 17 subjects. To each subject, the reference (desired signal) as well as the reproduced signals at the center of each zone were played back using Beyerdynamic DT 990 PRO headphones. The reference was always played back first, but the reproduced signals were played back in a random order, and we compared PM and P-VAST. The test was performed in a silent room, and four different scenarios of two bright zones were tested and repeated two times to each subject. Thus, each subject gave 16 answers. The audio examples simulated situations in which the two bright zones both contained music, contained music and speech, and both contained speech (available at `https://tinyurl.com/pvast2019`). We used $M = L = 8$, and the remaining parameters are summarized in Table 2 which also shows the results. Set 2 in Table 2 showed the least difference in preference between the two methods. We believe that this is due to that distortion is much more important than TIR when speech is the target and music is the interference. The opposite is the case for the mixed music case (set 3) where TIR is much more important. Although μ and V for each scenario were chosen empirically (since all scenarios have different characteristics), P-VAST clearly outperformed PM for all test signals.

5. CONCLUSION

We have proposed a framework called P-VAST for creating perceptually optimized sound zones. More concretely, we have shown how the reproduction error can be shaped according to the human auditory system. This has been done by computing masking thresholds from the desired signals in each zone. Via a simple proof-of-concept experiment and a listening test, we have shown that P-VAST significantly outperforms ACC and PM in terms of perceptual metrics.

6. REFERENCES

[1] F. M. Heuchel, D. Caviedes Nozal, F. T. Agerkvist, and J. Brunskog, "Sound field control for reduction of noise from outdoor concerts," in *Proc. 145th Conv. Audio Eng. Soc.*, New York, NY, USA, 2018.

[2] J. Cheer, S. J. Elliott, Y. Kim, and J.-W. Choi, "Practical implementation of personal audio in a mobile device," *J. Audio Eng. Soc.*, vol. 61, no. 5, pp. 290–300, 2013.

[3] X. Liao, J. Cheer, S. J. Elliott, and S. Zheng, "Design array of loudspeakers for personal audio system in a car cabin," in *Proc. 23rd Int. Congr. Sound Vib.*, Athens, Greece, 2016.

[4] J.-H. Chang, C.-H. Lee, J.-Y. Park, and Y.-H. Kim, "A realization of sound focused personal audio system using acoustic contrast control," *J. Acoust. Soc. Am.*, vol. 125, no. 4, pp. 2091–2097, 2009.

[5] M. F. S. Gálvez, S. J. Elliott, and J. Cheer, "Personal audio loudspeaker array as a complementary TV sound system for the hard of hearing," *IEICE Trans. Fundamentals*, vol. E97-A, no. 9, pp. 1824–1831, 2014.

[6] J.-M. Lee, T.-W. Lee, J.-Y. Park, and Y.-H. Kim, "Generation of a private listening zone; Acoustic parasol," in *20th Int. Congr. Acoust.*, Sydney, NSW, Australia, 2010.

[7] J.-W. Choi and Y.-H. Kim, "Generation of an acoustically bright zone with an illuminated region using multiple sources," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1695–1700, 2002.

[8] M. A. Poletti, "An investigation of 2D multizone surround sound systems," in *Proc. 125th Conv. Audio Eng. Soc.*, San Francisco, CA, USA, 2008.

[9] J. Francombe, P. Coleman, M. Olik, K. R. Baykaner, P. J. B. Jackson, R. Mason, S. Bech, M. Dewhirst, J. A. Pedersen, and M. Dewhirst, "Perceptually optimized loudspeaker selection for the creation of personal sound zones," in *Proc. 52nd Int. Conf. Audio Eng. Soc.*, Guildford, UK, 2013, pp. 1–9.

[10] T. Lee, J. K. Nielsen, J. R. Jensen, and M. G. Christensen, "A unified approach to generating sound zones using variable span linear filters," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Calgary, AL, Canada, 2018, pp. 491–495.

[11] J. Rämö, L. Christensen, S. Bech, and S. Jensen, "Validating a perceptual distraction model using a personal two-zone sound system," in *Proc. Meet. Acoust.*, Boston, MA, USA, 2017, vol. 30, p. 050003.

[12] J. Francombe, R. Mason, M. Dewhirst, and S. Bech, "Elicitation of attributes for the evaluation of audio-on-audio interference," *J. Acoust. Soc. Am.*, vol. 136, no. 5, pp. 2630–2641, 2014.

[13] M. B. Møller and M. Olsen, "Sound zones: On performance prediction of contrast control methods," in *Proc. AES Int. Conf.*, Guildford, UK, 2016.

[14] M. F. S. Gálvez, S. J. Elliott, and J. Cheer, "Time domain optimization of filters used in a loudspeaker array for personal audio," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 11, pp. 1869–1878, 2015.

[15] J. Donley, C. Ritz, and W. B. Kleijn, "Multizone soundfield reproduction with privacy- and quality-based speech masking filters," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 6, pp. 1041–1055, 2018.

[16] S. J. Elliott, J. Cheer, H. Murfet, and K. R. Holland, "Minimally radiating sources for personal audio," *J. Acoust. Soc. Am.*, vol. 128, no. 4, pp. 1721–8, 2010.

[17] T. Betlehem, W. Zhang, M. A. Poletti, and T. D. Abhayapala, "Personal sound zones: Delivering interface-free audio to multiple listeners," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 81–91, 2015.

[18] F. Olivieri, F. M. Fazi, S. Fontana, D. Menzies, and P. A. Nelson, "Generation of private sound with a circular loudspeaker array and the weighted pressure matching method," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 8, pp. 1579–1591, 2017.

[19] J.-H. Chang and F. Jacobsen, "Sound field control with a circular double-layer array of loudspeakers," *J. Acoust. Soc. Am.*, vol. 131, no. 6, pp. 4518–4525, 2012.

[20] K. Brandenburg and T. Sporer, "NMR and masking flag: Evaluation of quality using perceptual criteria," in *Proc. 11th Int. Conf. Audio Eng. Soc*, Portland, OR, USA, 1992.

[21] V. K. Kool and R. Agrawal, *Psychology of technology*, Springer International Publishing, 2016.

[22] G. H. Golub and C. F. Van Loan, *Matrix computations*, The Johns Hopkins University Press, 4th edition, 2012.

[23] B. C. J. Moore, *An introduction to the psychology of hearing*, Brill, 6th edition, 2013.

[24] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. H. Jensen, "A perceptual model for sinusoidal audio coding based on spectral integration," *EURASIP J. on Advances in Signal Processing*, vol. 2005, no. 9, pp. 1292–1304, 2005.

[25] H. Vogel, "A better way to construct the sunflower head," *Math. Biosci.*, vol. 44, no. 3-4, pp. 179–189, 1979.

[26] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[27] ITU, "ITU-T recommendation P. 862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," International Telecommunication Union, 2001.